Aalto University
School of Science
Master's Programme in Computer, Communication and Information Sciences

Andrei Kazlouski

Text style imitation to prevent author identification and profiling

Master's Thesis
Espoo, April 20, 2019

Supervisor: Professor N. Asokan, Aalto University

Advisor: Tommi Gröndahl

Aalto University
School of Science
Master's Programme in Computer, Communication and Information Sciences

ABSTRACT OF MASTER'S THESIS

Author: Andrei Kazlouski

Title:
Text style imitation to prevent author identification and profiling

Date: April 20, 2019 **Pages:** 39

Major: Security and Cloud Computing **Code:** SCI3084

Supervisor: Professor N. Asokan

Advisor: Tommi Gröndahl

Imitating the writing style of another author constitutes a tool to protect the privacy of the text author, while also can be used as an impersonation attack against the targeted person. At present, state-of-the-art deep learning methods have claimed success in both imitation of the targeted author and semantic retainment of the original text. By testing three representative text style imitation models on four varying datasets, I demonstrate that the methods are able to produce semantically correct transformations in only at most 50% of the transformed sentences. Furthermore, I demonstrate that the models are not able to consistently deceive the state-of-the-art LSTM and CNN deep learning classifiers for authorship classification. Combination of these two findings shows the studied models not to be applicable for real-life use cases. By studying the drawbacks of existing style imitation models, I reflect on ways of incorporating deep learning methods with other techniques to develop an imitation model that can be used for real-world application.

Keywords: deanonymization, stylometry, author identification, deception, text obfuscation

Language: English

Contents

Abbreviations and Acronyms

AI	Artificial Intelligence
CNN	Convolutional Neural Network
CAE	Cross-aligned Autoencoder Approach
SVM	Support-vector Machine
BT	Style Transfer Through Back Translation
A^4NT	Author Attribute Anonymity by Adversarial Training of Neural Machine Translation
LSTM	Long-Short Term Memory
NN	Neural Network
NLP	Natural language Processing
NMT	Neural Machine Translation
MT	Machine Translation
GAN	Generative Adversarial Network
ILT	Iterative Language Translation
RNN	Recurrent Neural Network

Contributions

This thesis presents a comparative analysis of the state-of-the-art text style imitation methods. The author and his thesis advisor Tommi Gröndahl conducted manual evaluation of the transformed sentences. Furthermore, Tommi Gröndahl helped with the designing comparative tests and choosing hyperparameters for training the models. He also contributed to the writing the code for the authorship classification of the transformed sentences. Tommi Gröndahl and Mika Juuti contributed to the debugging of the original code[1] for the A^{4}NT model.

[1] `https://github.com/rakshithShetty/A4NT-author-masking`

Chapter 1

Introduction

Recent advances in artificial intelligence (AI) techniques are a real threat to authors' privacy and anonymity [9] . A *Deanonymization attack* is uncovering the identity of an anonymous author. At present, author deanonymization techniques use technologies such as neural networks(NN) and statistical pattern recognition are able to disclose a number of privacy related author attributes such as gender, age, and sometimes even the identity of the text's author [6, 25, 27]. Furthermore, modern deanonymization methods are applicable to identify many types of the text authors: internet bloggers, book writers, journalists, and even government officials [3, 25, 32]. Naturally, the broad application of profiling can substantially restrict freedom of speech and right to stay anonymous.

Examples of author deanonymization can be found in real life. In 2011 a self-proclaimed Syrian female blogger was revealed to be an American male [8]. In 2013, Peter Millican and Patrick Juola used authorship attribution tools to identify J.K Rowling as the real author of *A Cuckoo's Calling* [16]. Juola also observes a court case of an asylym-seeker claiming to be oppressed for writing articles criticising his government. To support this claim he presented his other verified articles, and the court's decision was based on stylometry analysis of contested articles. More recent example shows that even well-known politicians can be targeted by *style imitation* attacks. In 2018 the vice president of the USA Mike Pence was presumably imitated in an anonymous New York Times Op-Ed criticising White House policy [3]. With the steady increase in the size and availability of targeted corpora and computing resources, the deanonymization attack will likely become even bigger privacy threat.

Style transformation can be used to mitigate the deanonymization attacks. Prior studies mostly used iterative language translation (ILT) [20] for these purposes. ILT is based on an assumption that translation to another language retains semantic properties of the original text, but loses some author-specific attributes. Many papers have studied how well translation across the intermediate language back to English can prevent deanonymization [5, 9, 20]. However, the ILT approach is able to perform only *obfuscation* of the original sentences at best and is not applicable for proper style imitation. For 2 author (A and B) settings we define *obfuscation* of a document as dropping the authorship classification accuracy close to random chance (50%), and *imitation* as classifying document as a different author (document by author A is classified as author B).

At present, algorithms can successfully perform image style imitation using modern deep learning techniques [13]. Gatys et al. employed Convolutional Neural Networks (CNNs) to render the semantic content of an image in different styles. A number of prior NLP studies used similar methods of perturbation injection for deceiving the authorship classification [18, 31]. These models fooled the state-of-the-art authorship classifier by incorporating

"noise" to the input text by addition, removal or replacement some of the signature words. The same classifier was also deceived by adding unrelated sentences to the text's topic. However, it is very likely that the semantic content would be altered by using any of those approaches. Recently, there have been attempts to adjust other deep learning techniques for automatic text style imitation. Representative models use stat-of-the-art natural language processing(NLP) methods like neural machine translations (NMTs) and generative adversarial networks (GANs).

It is unclear, whether these models can be used as a defensive tool against deanonymization attack. It is also unclear whether these models are suitable for real-life usage.

In this thesis, we compare existing deep learning methods of style imitation. To asses these imitation models, we experiment on two transformation types: gender imitation, and individual author imitation. To evaluate the experiment results, all models are tested on two tasks: *imitation success* and *semantic retainment*.

In summary, this thesis addresses three major questions:

Q1 Can the style imitation models successfully perform text style imitation?

Q2 How well can the style imitation models preserve original semantic content?

Q3 How do complexity, and grammar of corpora affect the imitation success?

Chapter 2

Background

Text style imitation can be used a defensive tool against deanonymization attack. This chapter discusses prior papers related to this topic.

2.1 Deanonymization as a privacy threat

Stylometry is the process of identifying authors based on the writing style. Stylometry has many applications in the security field. In recent years law enforcement have successfully employed stylometry approaches [10, 23] to identify criminals. Author identification also can provide assistance in fighting with cyberbullying. Stylometry methods were successfully applied to disclose identities of doppelgängers, i.e. users with multiple accounts. Galn-Garcia et al. [12] were able to link troll accounts to their real owners. Furthermore, Solorio et al. [39] detected doppelgängers on Wikipedia with 68% accuracy using support-vector machines (SVMs) with 239 linguistic features. Doppelgänger detection was also successfully performed by Afroz et al. [4]

However, fighting criminals is not the only utilization of author identification techniques. Attackers can use them to deanonymize text authors against their will. Author deanonymization as a privacy concern has been studied in a number of papers [5, 9, 15, 25]. In this section, representative studies are presented and described.

Adversarial purpose of deanonymization could potentially be bullying or harassment [5]. Moreover, another adversarial scenario was proposed by Brennan et al. [9]. They identify Alice as the anonymous complainer, and Bob as the abusive employer, where Bob is using stylometry to disclose the identity of Alice.

Narayanan et al. [25] presented a classifier that was able to correctly recognize an anonymous author from a dataset from 100,000 authors in over 20% of cases, making probability of guessing the correct author significantly higher than random guessing. Even though this experiment cannot be considered completely successful, because the percentage of identifying the author incorrectly is much bigger, the method can potentially be applicable for real-life usage. Tempestt et al. [26] shows that stylometry-based deanonymization attacks are a realistic privacy concern for small sets of authors($\leq$20).

By analyzing existing methods in author identification in their survey paper Gröndahl and Asokan [15] predict the deanonymization attack to be a growing privacy threat.

2.2 Existing style imitation models

In this section we describe three recent papers on style imitation. All three of them use deep learning methods like NMT and GANs. Current deep learning style imitation methods date back to 2017 and 2018. NMT methods are represented by Shen et al.'s [35] the cross-aligned autoencoder approach(CAE) and Prabhumoye et al.'s style-transfer through back translation(BT) method [29]. Finally, Shetty et al.[36] present an Author Attribute Anonymity by Adversarial Training of Neural Machine Translation (A^4NT) - the GANs-based approach to perform style imitation. Representative Style imitation models reviewed in this thesis are trained on non-parallel data, i.e., models do not have a list of mappings from an input to an output. The models address this issue by pairing style distributions instead of pairing individual words(sentences). This process is essentially guiding sentence generation.

2.2.1 Cross-aligned autoencoder approach (CAE)

The CAE approach is based on generating an intermediate style-independent representation of the target sentence. Combining this representation with a style-specific decoder model renders target sentences. First, the encoder generates a latent style independent content variable. During the generation process of the targeted style datapoint the decoder conditions on this variable. A process called *cross-aligned autoencoding* is used to align the encoding distributions between the source and target styles.

To assess semantic retainment of the transformed corpora, Shen et al. evaluated sentence transfer quality in a comparative manner, where the user study participants compared two transformations of a source sentence, and decided which one is better. The CAE paper also provides "fluency" rating of transformed sentences. However, fluency is not directly an evaluation of the semantic preservation, but an assessment of readability of the sentence. Fluency was rated from 1 (unreadable) to 4 (perfect). In their paper Shen et al. report following results for imitation success and semantic retainment:

- sentiment accuracy of transformed sentences - 78.4 %
- fluency rating of transformed sentences - 2.8

2.2.2 Style transfer through back translation (BT)

The BT approach is also based on generating an intermediate style-independent representation of target sentence. This method shares similarities with ILT and is based on the assumption that translation to another language will remove many style-specific features [29]. Prabhumoye et al. use pre-trained machine translation modules between English and French and vice-versa. First, they translate the original sentence to French. They then encode the acquired sentence with encoder part of French-to-English translator to lose author's specific traits. The model has two different decoders that are receiving the output of French encoding as an input. Those decoders are trained to generate sentences in distinct writing styles

In their paper Prabhumoye et al. performed gender transformation using restaurant reviews from Yelp[1]. They also trained the CAE model on the Yelp dataset to perform gender imitation. To assess semantic retainment of the transformed corpora, Prabhumoye et al. mimicked tests conducted by Shen et al. They also present comparative transfer

[1]https://yelp.com

quality, and fluency rating of BT and CAE gender transformation. User study participants preferred BT transformations over CAE transformations. In their paper Prabhumoye et al. report following comparative(**BT** and **CAE**) results for imitation success and semantic retainment:

- accuracy of the gender transformation - **BT**: 57.04%, **CAE**: 60.40%
- fluency rating of transformed sentences - **BT**: 2.81, **CAE**: 2.42
- preference score of transformed sentences - **BT**: 47.27%, **CAE**: 12.5%, **no pref.**: 41.36%

2.2.3 Author Attribute Anonymity by Adversarial Training of Neural Machine Translation (A^{4}NT)

A^{4}NT is a GAN-based approach to style transformation. A GAN consists of two separate neural networks - a *generator* and a *discriminator*. The generator is used to produce real-looking outputs and the discriminator's job is to identify fake ones. Both networks are in continuous competition as the generator tries to fool the discriminator, while the discriminator tries not to be fooled. A^{4}NT is an unsupervised model i.e. no parallel corpora are required. To trick a word-based LSTM author classifier, Shetty et al. train a GAN. A^{4}NT tries to preserve semantic content of the original sentence by maximizing *cyclic reconstruction probability*. A^{4}NT defines *cyclic reconstruction probability* as the probability of reconstruction the original sentence, when applying reverse style-transfer on transformed sentence [36]. For convenience, henceforth we refer to *cyclic reconstruction probability* as *cycle loss*.

Shetty et al. adopted the collections of blogs and political speeches for their experiments. To assess the semantic preservation of the transformed sentences, they conducted a comparative evaluation of the A^{4}NT model and the GoogleMT baseline [2]. They also report scaled evaluation of the transformed corpora. They rated sentences from 0 (the input and output sentences are not semantically related) to 5 (the input and output are equivalent in meaning). In their paper Shetty et al. report following results for imitation success and semantic retainment:

- F1 score of the blog transformation - 0.53 $\rightarrow$ **0.41**
- F1 score of the political transformation - 0.60 $\rightarrow$ **0.11**
- comparative evaluation of transformed sentences - **A^{4}NT**: 59.46%, **GoogleMT**: 43.76%
- scaled evaluation of transformed sentences - **A^{4}NT**: 4.51, **GoogleMT**: 4.16

2.2.4 Transformation examples of the imitation models

All these models are claiming high result in the accuracy classification of style transferred sentences. However, by analyzing the respective papers' examples [29, 35, 36] in table 2.1, a conclusion can be drawn that semantic retainment for a human reader is actually relatively low.

[2] https://translate.google.com/

While, CAE performs arguably successful sentiment transformation [35], it does alter semantic sense of the original sentence. BT's political style transformation [29] could potentially deceive the classifier, but obviously fails to adequately preserve semantic content of the original sentence. A^{4}NT's age transformation [36] form not only significantly alters its semantic content, but also adds a negation ("do care" vs "don't care").

Model	**Original sentence**	**Transformed sentence**	**category**
CAE	it was super dry and had a weird taste to the entire slice.	it was super flavorful and had a nice texture of the whole side.	sentiment
BT	i thank you, sen. visclosky.	i'm praying for you sir.	political
A^{4}NT	i'm kind of the term PERSON because i do care.	i'm tired of the system of PERSON PERSON because they don't care.	political

Table 2.1: Transformation examples from style imitation papers

The end goal of style imitation is to preserve the original meaning of the sentence, and fool the authorship classifiers to classify the transformed sentence as a different author. This can be achieved by applying paraphrasing, changing punctuation, introducing target specific text constructions like interjections, smiles, typos, etc. However, transformation examples of representative models suggest that existing style imitations models perform text generation conditioned on the targeted author writing style without proper paraphrasing of the original sentence. As a result, proper semantic retainment is not achieved. Hence, the real-life applicability of these methods is highly questionable.

2.3 Evaluation of machine translation

We use Machine Translation metrics to asses the text style imitation. Assessment of Machine Translation can be performed by both humans and automatic textual similarity measures. Papineni et al. presented the BLEU score [28] for evaluating the quality of text translation. BLEU first compute the n-gram matches between source and target sentences. After that, the algorithm adds the n-gram counts for the candidate sentence and divides by the number of candidate n-grams in the test corpus to compute a modified precision score, p_n, for the entire test corpus. Papineni et al. experimentally established n-gram order of 4 to have the biggest similarity with monolingual human judgements [28]. The METEOR score was introduced by Denkowski and Lavie [11], also for evaluating machine translation. The METEOR assessment also relies on n-gram overlap between source and target. However, METEOR also takes into account 4 other possible matches: exact match, stem match, synonym (based on WordNet lexical database[3]), and paraphrase (taking into account paraphrase tables, i.e., many-to-many matches).

[3]https://wordnet.princeton.edu/

Chapter 3

Problem description

Representative style imitation models employs deep learning classifiers to evaluate probability of the sentence been written by a particular author. We refer to such classifiers as *authorship classifiers*.

According to the authors of the state-of-the-art style imitation models [29, 35, 36], representative methods are able to perform proper author imitation and preserve the meaning of the original text. However, for assessing semantic retainment of the original corpora representative papers evaluated either fluency of transformed sentences or presented a comparison analysis of alternative approaches. Such tests do not comprise an appropriate evaluation scheme, considering that a transformed sentence can be perfectly readable, but have a different semantic content. Albeit Prabhumoye et al.[29] provided comparison results between the BT and CAE models, no prior studies conducted systematic comparison between state-of-the-art imitation models across many datasets of different corpora size, sentence length, sentence complexity, and grammar properties. Moreover, these studies do not classify the errors occurring during the imitation.

In regards to author imitation evaluation, all studied imitation models use only a single authorship classifier to assess imitation success. Moreover, the models do not report the classification results for transformations in both directions

In response to this problem, this thesis aims to investigate several ways of comparing all three state-of-the-art models, and make these comparisons objective. To adequately and fairly assess the performance of the state-of-the-art style imitation models, we plan to carry out extensive tests of representative models on the imitation success and deceiving the authorship classifier. To improve previous methods of assessing imitation success, we plan to employ an ensemble of differently trained state-of-the-art deep learning classifiers. Comparing imitation success of the models across different classifiers is likely to enhance the comparative evaluation of representative models. We also plan to compare automatic semantic retainment measures and to conduct a manual evaluation of error types occurring in transformed corpora.

Below, comparison tests that we are planning to introduce are summarized:

1. **Imitation** evaluation with an ensemble of authorship classifiers
2. **Automatic** evaluation of semantic retainment.
3. **Manual** evaluation of semantic retainment.

Chapter 4

Comparative evaluation

4.1 Environment

This section describes our training setup and details of trained models. Replicated deep learning models are using the PyTorch [1] and TensorFlow [2] frameworks. We converted all the models to Python 3.6 for our experiments. For training the models we used computer with 4 GB GPU memory. When pre-trained models were available, we used pre-provided models by Shetty et al. and Prabhumoye et al. Links for all pre-provided models are presented in Appendix A.

4.2 Experiment Data

Table 1 describes properties of the datasets used for our experiment. We are going to describe each dataset separately below.

Dataset	**Sentence length (words)**				**sentences**	**vocab size**
	Mean	**Std.**	**Min.**	**Max.**		
Yelp	18.16	10.80	4	101	3,208,136	19,996
Blog	12.76	7.39	2	30	3,379,779	264,966
Individual author	11.19	8.80	1	110	31,319	16,498
Political speech	18.57	13.25	2	294	65,485	16,348

Table 4.1: Properties of the datasets.

4.2.1 Yelp gender data (YG)

The dataset consists of restaurant reviews. Since most reviews are relatively short, and their purpose is to give an assessment to a restaurant, they tend to consist of grammatically correct and short sentences mostly using common English vocabulary and rarely typos. Later statement is illustrated in Table 4.1. Despite the large size of the Yelp corpora, the vocabulary size is relatively small.

Originally this dataset was collected by Reddy and Knight's [30] from a corpus of reviews from the Yelp Dataset Challenge 2016. Reddy and Knight establish **gender** attribute of the reviewers by analyzing first names.

4.2.2 Blog gender data (BG)

Unlike YG blogs are often written in a "colloquial style". While the purpose of restaurant reviews is to give a clear assessment, blogs include more general conversational content and can be used for wider range of purposes. As a result, the blog sentences tend to be less informative and grammatically imperfect. Also often blogs consist of very short sentences. Furthermore, the blog dataset contains more typos, abbreviations such as LOL or OMG, and interjections like "ouch" or "sigh".

Although BG and YG datsets are similar sizewise, Table 4.1 illustrates that the drastic difference in vocabulary sizes. Due to frequent occurrence of uncommon and wrongly spelled words in the blog corpora, the blog vocabulary size is 13 times bigger than YG one.

Originally this dataset was collected by Schler et al. [32]. The dataset was collected from collections of blogs at blogger.com. It contains 19,320 documents. Every document is a single blog post, consisting of many sentences by a single author, and has been annotated for the author's *identity*, *gender*, and *age*. Shetty et al. [36] used this dataset for training the gender and age models for A^4NT. They used the Stanford CoreNLP tool [22] to segment the documents into sentences. They split the entire dataset of 19,320 documents into 13,636 training documents, 2.885 test documents and 2,799 validation documents.

4.2.3 Individual author data (AB)

The individual author data is derived from the blog dataset. We extracted 2 authors with the largest amount of text. One of the authors is a middle aged male, and other is a young female, making these 2 authors distinct not only in identity but gender and age as well. For convenience, we name the authors *Bob* and *Alice* respectively. The central idea of doing individual author transformation is to study the effect of limited training data on quality of style transformation. Similar to BG, AB sentences are also of the colloquial nature.

Shetty et al. [36] used only sentences that do not exceed 30 words for BG. To increase the size of the individual author corpora, we used all available sentences for 2 authors. Hence, the length properties of AB dataset differs from the the blog dataset. This can be observed in Table 4.1

4.2.4 Political speech data (TO)

The dataset consists of political speeches of two American Presidents Barack Obama and Donald Trump. Since both of them are talking about similar topics it is feasible to perform style imitation, and compare the results. Applying style transformation on the corpora of such acclaimed Presidents is especially compelling since real-life imitation of major politicians happened before [3]. Since most of the Presidents' speeches are prepared by experienced speechwriters they tend to be grammatically correct and concise. Shetty et al. collected this dataset from the American Presidency Project [42]. They split the data in 372 documents with each document being a separate speech by an author. To avoid authorship classifiers relying on names of specific people or places, Shetty et al. altered names of all named entities like organizations, institutions, and personal names etc. Total size of the dataset is 65,485 sentences. Table 4.1 shows that the political dataset is the most diverse in regard to sentences' length, the shortest sentence consists of only 2 words, and the longest one of 294.

4.3 Comparative empirical evaluation of style imitation techniques

In this section we describe the procedure of our experiments and details of the tests we conducted on the transformed data to test the reliability of semantic retainment and the success of style imitation.

All imitation models are trained to perform style imitation for every dataset. We apply the same set of tests on each corpora transformation. Since both CAE and A^{4}NT truncate output sentences at 20 words, for every dataset we created a subset of the test set of sentences below 21 words. We did the classifications tests on both original and reduced test sets. Since potential proper paraphrase is likely to be of the same size as the original sentence, we conducted automatic textual overlap and manual evaluation tests only on the truncated subsets to adequately assess the semantic retainment. It is also evident that truncating a big sentence to 20 words is very likely to alter its original meaning. We transferred the style of the test sentences and then tested the classification accuracy of the generated sentences for the same label. For example, to test the success of the imitation of the style transferring from male to female, we tested generated sentences for the male label.

In Introduction we defined *imitation* for 2 authors settings. However, due to the authorship classification accuracy being a continuous measure, it is not evident how far below 50% classification should be reduced to achieve imitation. To address this issue we will consider imitation successful if the classification drops from above 50% to below 50%. The question, then, arises: is reducing authorship classification accuracy from 51% to 48% an imitation? There is no a clear answer, but we will address every such marginal case individually.

4.3.1 Imitation evaluation

To fairly assess the imitation success of the tested style transformation models we tested the transformed test datasets on a number of authorship classifiers. Authorship classifiers used for this experiments are state-of-the-art deep learning classifiers. The BT method trains a CNN classifier. The A^{4}NT model uses 2 independent Long-Short Term Memory (LSTM) classifiers throughout the training process. "LSTM" is the baseline classifier used for the initialization of a discriminator part of the GAN, " LSTM_{GAN} " is the authorship classifier that is a result of training the GAN.

LSTM and CNN. A LSTM and CNN classifiers are the authorship classifiers that are trained to classify the original corpora. These classifiers represent a baseline classification of the imitation success. Deceiving these baseline authorship classifiers can be considered as proper imitation of the targeted author.

LSTM_{GAN}. A LSTM_{GAN} classifier is trained to recognize imitation attempts by the generator and, therefore, should likely classify transformed sentences closer to the true label. Naturally, " LSTM_{GAN} " is trained to recognize imitaion attempts by A^{4}NT. However, it is compelling to study whether BT and CAE can fool " LSTM_{GAN} ".

4.3.2 Assessment of semantic retainment

Automatic evaluation of semantic retainment. To assess the textual overlap of the tested models we calculated the BLEU and METEOR score using original sentences as the reference and transformed sentences as the test.

It it possible to achieve perfect semantic retainment by not changing the original sentence at all. Naturally the combination of identical and completely changed sentences can

yield a relatively high BLEU/METEOR score. Hence, we count the ratio of such sentences(*Same*), we also separately calculate the BLEU and METEOR score for the transformed sentences(*BLEU non-id*, *METEOR non-id*) that are different from the original sentences. We also noticed that A^{4}NT often removes word/words without replacing them with any paraphrase. This can likely result in semantic loss. However, this fact can be concealed by high METEOR or BLEU due to high n-gramm overlap. Therefore, we measure the percentage of the transformed sentences the vocabularies of which are a proper subset of the original sentences (*Subset*).

We tested whether removing punctuation affects BLEU and METEOR. Since punctuation does not drastically affect the human evaluation of the semantic content of the text, it can be considered as the "noise" one can add to deceive the classifier and retain semantics. However, measurement changes were very slight and, therefore, we did not remove punctuation in the tests reported here.

Manual evaluation of semantic retainment. To estimate the possibility of this models to be used for any real-life transformation, we conducted a small-scale manual evaluation consisting of evaluating 50 sentences from each transformation for every model. For each sentence we introduce 6 possible assessments:

- S - exact(same) match of the original sentence
- P - perfect paraphrase of original sentence
- R - reasonable paraphrase of the original sense. I.e semantic content is very close to the original, and the sentence does not contain errors and perfectly understandable for a human reader.
- W - inappropriate word change
- O - inappropriate omission of words from the original sentence
- A - inappropriate addition of words to the original sentence
- G - presence of grammatical errors in the transformed sentence

Labels *P*, *R* or *S* mean that the transformation is successful and is transparent for a human reader, while other labels mean that the transformation failed. A sentence can be marked with only one successful label, but any number of failed labels e.g. *WOG* or *WA*

4.4 Results

In this section we describe obtained results for every studied model across 4 datasets. Since transformed test sentences are tested against the original label, the lower classification results are the better. There is no uniformly "good" result for automatic MT metrics. Therefore, we consider the BLEU and METEOR scores above 0.5(50%) to be good. Since each manual evaluation consists of only 100 sentences, we address each of them individually.

4.4.1 Blog Gender (BG)

The test results for BG imitation are presented below.

Imitation success. Table 4.2 illustrates the results for the classification of the gender imitation in BG. It is clear that all authorship classifiers are always leaning towards the female

class, and seldom can correctly predict the male class for original test data. This is potentially due to the imbalance of the training dataset. The female training set has 1,309,137 sentences, while male training set has only 1,044,188 sentences. The results suggest that the blog data does not have many distinctive features between male and female. Therefore, the training corpora size is the most prevalent feature for classification in BG. The LSTM_{GAN} classifier is the only classifier that can predict the true label with a probability higher than 50%. Regarding imitation success, we can conclude that none of the models are able to perform proper male imitation. Classification accuracy of the transformed female corpora is not dropping below 50%, but actually rises. Even though classification accuracy of transformed male corpora drops, it is hardly an imitation, because none of the baseline classifiers are able to classify original male sentences better than a random chance. It is worth noting that LSTM_{GAN} classifier can not only recognize imitation attempts by A^{4}NT, but also perform the most balanced classification of other models' transformations.

Classifier	**Original**		**A^{4}NT**		**BT**		CAE	
	f	m	f → m	m → f	f → m	m → f	f → m	m → f
CNN	0.78	0.38	0.8	0.34	0.78	0.22	0.79	0.25
LSTM	0.74	0.39	0.74	0.37	0.79	0.21	0.8	0.24
LSTM_{GAN}	0.6	0.54	0.59	0.54	0.56	0.40	0.59	0.47

Table 4.2: Classification measures of transformed sentences in BG.

Regarding classification of truncated BG, Table 4.3 indicates that the results for truncated corpora are similar to the full dataset.

Classifier	**Original**		**A^{4}NT**		**BT**		**CAE**	
	f	m	f → m	m → f	f → m	m → f	f → m	m → f
CNN	0.78	0.38	0.82	0.31	0.8	0.2	0.8	0.24
LSTM	0.74	0.39	0.76	0.35	0.74	0.31	0.78	0.28
LSTM_{GAN}	0.6	0.54	0.59	0.53	0.54	0.42	0.58	0.49

Table 4.3: Classification measures of transformed sentences in truncated BG.

Semantic retainment. Textual similarity of the gender imitation of BG data is presented in Table 4.4. It is clear that A^{4}NT performs better than other models in retaining semantic content in gender transformation in BG. However, A^{4}NT replaces words only in less than 10 % of the sentences. The remaining sentences are either the same or a product of removing words. This approach might fool the authorship classifier, but is not a semantically appropriate imitation scheme. Meanwhile, the number of the same sentences is considerably lower for CAE, and BT introduces word replacement or addition in more than 90% of transformations. Furthermore, Table 4.4 shows that the METEOR score is higher than BLEU for CAE and BT. Those models are performing word replacements more frequently and the changes can occasionally match the synonym or paraphrase tables of the METEOR score. For A^{4}NT BLEU is higher than METEOR due to high n-gram overlap, since most sentences contain the same words as the original ones.

Table 4.5 illustrates the manual error analysis of transformed sentences in BG. It is clear, that all studied models can rarely produce a sentence without changing the semantic content of the original. In fact, only A^{4}NT can transform more than 10% of the sentences (labels *S*,*P* or *R*) without losing the original meaning of the text.

Table 4.6 illustrates examples of gender transformations in BG. It is evident that the most common error type is an inappropriate word replacement. Furthermore BT occasionally adds

Test	**A^{4}NT**		**BT**		**CAE**	
	f → m	m → f	f → m	m → f	f → m	m → f
Same	0.54	0.49	0.04	0.04	0.19	0.15
Subset	0.37	0.43	0.04	0.04	0.19	0.26
BLEU	81.41	78.35	18.56	16.95	29.49	26.93
BLEU non-id	65.25	63.24	17.67	16.19	25.5	24.18
METEOR	54.46	52.29	23.11	22.45	30.69	27.86
METEOR non-id	43.93	42.89	22.6	21.96	28.54	26.24

Table 4.4: Textual overlap measures of transformed sentences in BG.

	S	**P**	**R**	**W**	**A**	**O**	**G**
CAE	11	1	1	69	6	17	9
BT	5	0	1	92	20	5	6
A^{4}NT	43	5	3	8	1	20	2

Table 4.5: Manual evaluation of semantic retainment in BG (*total of 100 sentences*)

negations to affirmative sentences. As mentioned above, A^{4}NT examples outperform other imitation models by removing words in transformed sentences.

4.4.2 Yelp Gender (YG)

We trained the generator and the classifier of the A^{4}NT model on full training dataset, but due to hardware limitations (Appendix 4.1) we were not able to train the GAN itself on the entire training corpora. Instead we used a 100,000 sentences subset. A^{4}NT code allows to use different loss types to ensure that the generator part of the GAN can not apply changes too often. We used cycle loss for our experiments (Appendix A). We compared obtained results to not using any type of loss and got a clear distinction. Not using cycle loss results in major changes and, therefore, significant semantic changes.

The Yelp dataset contains the independent data subsets for training both a generative model and a classifier. Therefore, for the experiments on the Yelp data we trained the additional "LSTM^2" authorship classifier (a number 2 denotes second dataset used for training the classifier). This classifier is trained on the independent subset of training data and, therefore, is likely to be harder to deceive than a naive classifier. Hence, a LSTM^2 classifier can be considered a "stronger" version of LSTM^1 classifier.

Imitation success. In Table 4.7 results for the classification of the YG transformations are presented. Unlike in BG, all classifiers are able to predict the true labels with relatively high accuracy. It can be argued that YG has more gender distinctive features than BG, since all classifiers are able to classify both female and male sentences with relatively high probability. In regards to imitation success , both BT and A^{4}NT are able to decrease the classification accuracy of all baseline classifiers below 50% and as a result perform proper imitation of the targeted writing style. Meanwhile, CAE is not able to consistently perform imitation

Technique	Sentence	Label
Original	"sadly , however , this wretched illness also leaves me horribly disfigured ."	
CAE	"sadly , however , this has still taught me <unk> my brain <unk> ."	W
BT	"unfortunately , however , this site 's hopefully i 'm mellow ."	W
A^{4}NT	"sadly , however , this illness shelby also leaves me horribly ."	WO
Original	"if we were in school i would vote you for prom queen 3 ."	
CAE	"if we were in college i love we guys i miss music queen in <unk> ."	W
BT	"if we were in love i would n't be able to see you for prom ."	W
A^{4}NT	"if we were in school i would vote you for prom queen ."	P
Original	"seriously , this squirrel pooped so much it 's insane ."	
CAE	"seriously seriously , it 's my old it 's so insane ."	WA
BT	"seriously , this kid 's so much that it 's the folie ."	W
A^{4}NT	"seriously , this . com so much it 's insane ."	OW
Original	"i miss me my college parties ."	
CAE	"i miss my friends on my students ."	G
BT	"i ca n't remember my dad ."	W
A^{4}NT	"i miss me my college parties ."	S
Original	"i got new specs liao le le ELIP ."	
CAE	"i got new home le le le ELIP ."	W
BT	"i have n't been able to get some new caractéristiques the story ."	WA
A^{4}NT	"i got new . liao le le ."	W
Original	"yep , we are truely an internet couple ."	
CAE	"yep , we are truly an different two days ."	W
BT	"yes , we are really an interesting guy ."	WG
A^{4}NT	"yep , we are truely an internet couple ."	S

Table 4.6: Examples of sentence transformation in BG

across different classifiers[1] . Moreover, compared to other models, classification results of CAE are different for the A^{4}NT classifier trained on separate training data. LSTM^2 is able to recognize male to female imitation attempts and dropping the original accuracy only from 79% to 60%. Unlike BG, BT and CAE models yield similar classification results across all LSTM classifiers in YG transformations. Naturally, A^{4}NT cannot fool the LSTM$_{\text{GAN}}$ classifier.

Table 4.8 demonstrates that in YG the authorship classifiers can better recognize imitation attempts by A^{4}NT when the tested corpora is truncated. Similarly, BT results for truncated BG are slightly worse across the majority of the classifiers. Finally, It is hard to establish any correlation for the CAE model.

Semantic retainment. Table 4.9 indicates that automatic semantic retainment measures are similar for both BG and YG. A^{4}NT outperforms other models, but mostly due to the fact that changes it applies are relatively marginal. Similar to BG, the METEOR score exceeds BLEU for the BT and the CAE models. This is likely due to synonym matches of

[1]Shrimai Prabhumoye shared with me the CAE model, he used in his paper. To our surprise, the model seem to be highly overtrained Every adversarial sentence is just a repetition of a set of words. Therefore, we trained the CAE YG model ourselves. Example of the generated sentences with Prabhumoye CAE model can be seen in Appendix A

Classifier	**Original**		**A^{4}NT**		**BT**		**CAE**	
	f	m	f → m	m → f	f → m	m → f	f → m	m → f
CNN	0.78	0.85	0.42	0.42	0.4	0.44	0.54	0.46
LSTM^1	0.78	0.8	0.43	0.39	0.45	0.42	0.61	0.37
LSTM^2	0.82	0.79	0.47	0.39	0.47	0.38	0.38	0.6
LSTM$^1_{\text{GAN}}$	0.76	0.73	0.53	0.6	0.45	0.45	0.59	0.4

Table 4.7: Classification measures of transformed sentences in YG.

Classifier	**Original**		**A^{4}NT**		**BT**		**CAE**	
	f	m	f → m	m → f	f → m	m → f	f → m	m → f
CNN	0.78	0.85	0.55	0.49	0.48	0.41	0.52	0.49
LSTM^1	0.78	0.8	0.52	0.52	0.44	0.48	0.45	0.56
LSTM^2	0.82	0.79	0.56	0.51	0.49	0.41	0.57	0.44
LSTM$^1_{\text{GAN}}$	0.76	0.73	0.66	0.57	0.47	0.46	0.49	0.54

Table 4.8: Classification measures of transformed sentences in truncated YG.

METEOR.

Test	**A^{4}NT**		**BT**		**CAE**	
	f → m	m → f	f → m	m → f	f → m	m → f
Same	0.49	0.2	0.01	0.01	0.01	0.01
Subset	0.43	0.22	0.06	0.05	0.13	0.15
BLEU	78.35	45.85	15.92	15.06	8.73	8.67
BLEU non-id	63.29	37.26	15.63	14.82	8.24	8.15
METEOR	52.29	33.92	21.34	19.68	15.29	15.83
METEOR non-id	42.89	30.05	21.22	19.58	15.14	15.68

Table 4.9: Textual overlap measures of transformed sentences in YG.

Table 4.10 illustrates the error analysis of transformed sentences in YG. It is clear, that all studied models can rarely produce a sentence without changing the semantic content of the original. In fact, only A^{4}NT can transform more than 5% of the studied 100 sentences (labels *S*,*P* or *R*) without altering the original meaning of the text.

	S	**P**	**R**	**W**	**A**	**O**	**G**
CAE	0	1	1	95	6	34	9
BT	1	0	2	97	17	5	2
A^{4}NT	20	3	3	65	1	20	4

Table 4.10: Manual evaluation of semantic retainment in YG (*total of 100 sentences*)

Table 4.11 illustrates wide range of semantical errors that are present in YG transformations. Most frequent are inappropriate word changing. Furthermore BT occasionally changes affirmative sentences to negations, and vice-versa. CAE example transformations suggest that the model retains the least semantic information of the original sentence.

Technique	Sentence	Label
Original	“all perfect with my meal .”	
CAE	“all my friends for the meal .”	W
BT	“everything ’ s humongous with my wife .”	W
A^{4}NT	“all perfect with my meal .”	S
Original	“you name it , they offer it to you .”	
CAE	“you can be it , it ’ s not .”	W
BT	“mainly you get it , they ’ re going to you .”	W
A^{4}NT	“you name it , they offer it .”	R
Original	“as we are waiting , i notice a couple being seated that came in much after us .”	
CAE	“so we were seated at all , i had a few times .”	W
BT	“as we donnie , i ’ m a couple of which is looking in and out in whatevs .”	WG
A^{4}NT	“as we were waiting , i noticed a couple why not often that came in much .”	OGW
Original	“sure will go back again .”	
CAE	“will be going back again .”	P
BT	“maybe monkey .”	W
A^{4}NT	“will go back again .”	R
Original	“they gave a small tiny modern .”	
CAE	“they also a nice touch .”	W
BT	“they did n’t have a great time .”	W
A^{4}NT	“they gave a small cigarette smoke .”	W
Original	“i give that margarita 00 limes for its balanced yumminess .”	
CAE	“i felt like that i ’ re for 0 stars .”	WG
BT	“i guess that ’ s 00 stars for his food selection .”	W
A^{4}NT	“i give that 00 min jar for its remarkable .”	W
Original	“he really made us feel welcome and i will definitely be back again !”	
CAE	“he will definitely be back again and i will definitely be back again .”	W
BT	“he really did n’t have to say and i would be back again !”	WA
A^{4}NT	“he really made it key competition and that will definitely be back again .”	W
Original	“while i ’ m waiting for my things i only have a suitcase of clothing .”	
CAE	“i ’ m not a fan of that i ’ re a fan of them .”	GW
BT	“while i am looking for my wife , i ’ m only a lot of money .”	W
A^{4}NT	“while i ’ m waiting for my things only i have a series of spices.”	W

Table 4.11: Examples of sentence transformation in YG

4.4.3 Individual author (AB)

For training A^{4}NT on AB, we used cycle loss (AppendixA).

Imitation success. In Table 4.12 classification results of AB transformations are presented. It is evident that individual authors have more distinctive features than different genders. All authorship classifiers display very high classification accuracies of the original data. In regards to imitation success , both BT and CA are able to perform style imitation in both directions by decreasing the classification accuracy of all classifiers below 25% and below 20% with two exceptions. Meanwhile, the A^{4}NT model is more conservative and is not able to deceive any of the classifiers .

Classifier	**Original**		**A^{4}NT**		**BT**		**CAE**	
	Alice	Bob	A → B	B → A	A → B	B → A	A → B	B → A
CNN	0.89	0.91	0.78	0.6	0.07	0.13	0.22	0.09
LSTM	0.86	0.94	0.73	0.69	0.06	0.16	0.18	0.15
LSTM_{GAN}	0.91	0.91	0.82	0.67	0.07	0.12	0.23	0.09

Table 4.12: Classification measures of transformed sentences in AB.

Semantic retainment. Results of the textual similarity of gender imitation in AB data are presented in Table 4.13. It is evident that models perform massively worse in cases of limited data. Due to considerable number of unchanged sentences, only the A^{4}NT method produces results that exceed 10% in either of scores. Both CAE and BT shows very poor results in both METEOR and BLEU. For Bob imitation BT transformations contain 0 counts of 4-gram overlaps. Therefore the BLEU score evaluated to 0.

Test	**A^{4}NT**		**BT**		**CAE**	
	A → B	B → A	A → B	B → A	A → B	B → A
Same	0.24	0.15	0.001	0.02	0.02	0.01
Subset	0.24	0.19	0.05	0.07	0.02	0.01
BLEU	25.35	21.96	0	0.5	2.47	1.77
BLEU non-id	19.7	18.42	0	0.4	2.23	1.64
METEOR	28.17	25.92	6.18	6.32	7.77	6.26
METEOR non-id	24.82	23.85	6.17	6.13	7.39	6.15

Table 4.13: Textual overlap measures of transformed sentences in AB.

Table 4.14 illustrates the error composition of transformed sentences in AB. It is clear, that cases of semantic retainment are almost non-existent. The most frequent error category is **W** i.e the "wrong change". It appears in 91% of the studied transformations. For BT every single evaluated sentence contains **W**.

	S	**P**	**R**	**W**	**A**	**O**	**G**
CAE	0	0	1	99	1	20	0
BT	0	0	0	100	2	26	0
A^{4}NT	16	0	0	75	12	20	0

Table 4.14: Manual evaluation of semantic retainment in AB (*total of 100 sentences*)

Table 4.15 illustrates that produced sentences in most cases are completely unrelated to the orignal ones. The presence of french words in BT transformations and <unk>s in CAE sentences is likely a consequence of a small corpora size and vast amount of uncommon words.

4.4.4 Trump Obama (TO)

The test results for TO style imitation are presented below.

Technique	Sentence	Label
Original	“so was n’t too bad .”	
CAE	“so im going to see .”	W
BT	“it was going to get !”	W
A^{4}NT	“so was n’t too bad .”	S
Original	“i concluded i ’ll be drawing the 007 opening screen in every movie..”	
CAE	“i think i was going to <unk> the <unk> and i want to be a <unk> !”	W
BT	“i think i got to go at my l and then i ai”	WO
A^{4}NT	“i ’ll be able in the meantime i ’ll speed in”	WO
Original	“so after school now ... .”	
CAE	“so so much !”	W
BT	“so im going to go .”	W
A^{4}NT	“so after school now ... .”	S
Original	“wish me luck on not having a nervous breakdown !”	
CAE	“well after this year is a lot of <unk> ... .”	W
BT	“i was was pour , , function function lol .”	W
A^{4}NT	“wish me not even on a smart !”	WO
Original	“we have a n64 at work here , but mostly we just play mario kart .”	
CAE	“we had a <unk> <unk> <unk> , and then headed out of <unk> . ”	WO
BT	“i was fattoria nous , , , , , and implanter au .”	WO
A^{4}NT	“we have a at work , but we just wanted to play at times .”	WO
Original	“i think we just slept in ...”	
CAE	“i think i did to go ... ...”	W
BT	“i was was dans for permettra-t-elle ... .. ”	W
A^{4}NT	“i think we just got in london ...”	W

Table 4.15: Examples of sentence transformation in AB

Imitation success. In Table 4.16 classification results of TO transformations are presented. The length of training subsets for Obama and Trump are 10,430 and 31,662 sentences respectively. Due to the imbalance of the training dataset and authorship features, none of the baseline classifiers are able to predict class “Trump” for original data. The LSTM_{GAN} learns to detect imitation and is able to distinguish between Trump and Obama.
Semantic retainment. Regarding imitation , we can conclude that all models are able to drop the classification results for the LSTM_{GAN} classifier, and perform obfuscation of the targeted writing style. However, the LSTM and the CNN always predict “Obama” and, therefore, are not conveying any information for the TO transformations.

Classifier	**Original**		**A^{4}NT**		**BT**		**CAE**	
	Obama	Trump	O → T	T → O	O → T	T → O	O → T	T → O
CNN	1	0	1	0	1	0	1	0
LSTM	0.999	0.0002	0.99	0.0002	1	0	0.92	0.03
LSTM_{GAN}	0.71	0.7	0.54	0.57	0.42	0.57	0.32	0.54

Table 4.16: Classification measures of transformed sentences in TO.

Table 4.17 illustrates that in cases of limited data values of “Same sentences” and “Subset” fields are close to 0. That means that imitation models apply more changes, and as a

Test	**A^{4}NT**		**BT**		**CAE**	
	O → T	T → O	O → T	T → O	O → T	T → O
Same	0	0	0.003	0.02	0.01	0.02
Subset	0.08	0.16	0.01	0.02	0.32	0.11
BLEU	15.41	22.03	1.67	2.66	1.68	2.72
BLEU non-id	15.41	22.03	1.35	2.4	1.62	2.47
METEOR	20.31	23.68	8.83	10.1	6.94	8.76
METEOR non-id	20.31	23.68	8.8	9.98	6.86	8.53

Table 4.17: Textual overlap measures of transformed sentences in TO.

result automatic semantic retainment measures in case of limited data is considerably worse than BG or YG. Due to usage of the cycle loss A^{4}NT does less changes and outperforms other models. The BLEU and METEOR scores are very low for CAE and BT models.

Table 4.18 illustrates the error composition of transformed sentences in TO. It is clear, that cases of semantic retainment are almost non-existent. The most frequent error category is the "wrong change". It appears in 77% of the studied transformations.

	S	**P**	**R**	**W**	**A**	**O**	**G**
CAE	0	0	1	47	8	65	0
BT	0	1	1	98	7	18	10
A^{4}NT	1	0	3	87	18	12	3

Table 4.18: Manual evaluation of semantic retainment in TO (*total of 100 sentences*)

Table 4.19 illustrates that produced sentences in most cases are completely unrelated to the original ones. CAE example transformations shows that model often substitute entire sentence with a dot.

Technique	Sentence	Label
Original	“i still believe in you .”	
CAE	“i said you .”	OW
BT	“i ’ve going to be it .”	W
A^{4}NT	“i really still you very bad in”	WA
Original	“the question is what we do about it .”	
CAE	“it ’s going to happen .”	OW
BT	“that ’s going to be the country.”	W
A^{4}NT	“the media is what about it they do n’t know”	WA
Original	“the press does n’t like that term .”	
CAE	“this is not a lot of LOCATION .”	W
BT	“governor PERSON is not going to be .”	W
A^{4}NT	“the predicted does n’t like that future”	W
Original	“that will save you money at the pump .”	
CAE	“.”	O
BT	“it ’s going to be the country .”	W
A^{4}NT	“that will require at you like the money”	GW
Original	“thank you very much .”	
CAE	“thank you .”	R
BT	“thank you so much . .”	P
A^{4}NT	“thank you very much much”	A
Original	“gangs will disappear .”	
CAE	“.”	O
BT	“workers are n’t be .”	WG
A^{4}NT	“trust will disappear”	W

Table 4.19: Examples of sentence transformation in TO

Chapter 5

Discussion

In the beginning of the research, we wanted to study modern style imitation models and evaluate the retainment of semantic content for the original corpora, despite claims of the high retainment by authors of representative papers. However, the obtained results also proved that state-of-the-art model can not consistently perform proper style imitation across different datasets.

Imitation success. Deep learning techniques like LSTM and CNN are naturally good at teaching encoder-decoder networks to generate text sequences. Hence, it is not unexpected, that studied models are often able to fool the authorship classifiers. However, when the datasets are imbalanced (BG and TO), the representative models are not able to consistently perform proper imitation. Although the reason for this might be problems in classification of imbalanced data, some real-life use cases for style imitation might operate with datasets. Therefore, some mechanisms of addressing this issue must be implemented. Overall, the obtained results suggest that the state-of-the-art models were able to lower classification results for the true label of the transformed sentences for balanced datasets (YG and AB). Corpora size, text complexity, and grammar seem to have a large impact on the imitation success. YG sentences being simpler and grammatically correct outperformed BG transformations in the authorship classification compared to BG. Regarding imbalanced data, to our surprise, a LSTM_{GAN} classifier is the only classifier that can handle a problem of imbalanced training dataset. The CNN and LSTM classifiers favour the label with the most training data, whereas a LSTM_{GAN} seemingly relies on slightly different features different features. It is especially evident in case of TO. Baseline classifications with CNN and LSTM are **Obama**:100% and **Trump**:0%. LSTM_{GAN} is able to classify both labels with 0.7 probability. In case of BG LSTM_{GAN} is able to *even* the classifications results from **female**:0.77 and **male**:0.36 to **female**:0.59 and **male**:0.54. Although LSTM_{GAN} authorship classification results are close to a random chance, it is clear that during training the generator of the GAN, baseline classifier discloses different less transparent classification features.

Semantic retainment. Regarding semantic retainment, Chapter 3 illustrates that studied models are not able to reliably retain semantic content of the original sentences. None of the imitation models are able to retain semantic content of the original sentences for neither of studied datasets. Furthermore, colloquial style corpora is illustrated to produce worse transformations in cases of limited training data (TO outperformed AB, and YG outperformed BG). The size of training corpora is proven to be the most decisive factor for the imitation success of deep learning models. It is not unforeseen that reducing the size of the training dataset negatively impacts the performance of the studied models. Most transformation examples in TO and AB are completely altering the meaning of the original sentence. Manual evaluation in Chapter 4 indicates that shorter sentences more often preserve original

meaning. Majority of the sentences that was ranked with "S", 'P' or "R" are less than 10 words. Manual evaluation illustrates that the word substitutions that the models apply are semantically unjustified. Since short sentences are less likely to receive any changes, they are more likely to preserve the original meaning.

Overall, A^{4}NT outperforms BT and CAE in semantic retainmnet for every tested dataset transformation, due to the presence of loss penalty. BT and CAE have neither any penalty for changing the original sentence too much, nor any max edit distance to enforce semantic retainment. This leads to significant changes in the original sentence. It can be observed in cases of limited training corpora (AB and TO). CAE and BT are able to achieve almost perfect imitation by changing the original sentence to the set of words that are used exclusively by the target authors. Whereas, A^{4}NT produce much better sentences and drastically outperform other imitation models in semantic retainment task.

To generate adversarial transformations, BT and CAE are using an intermediate style-independent representation of the target sentence. The general idea is to generate so-called "normalized" state first, and then generate an output conditioned on the targeted writing style. BT and CAE try to decode a sentence in two distinct styles. However, the assumption that for every sentence different style-dependent representations exist might not be true. The question, then, arises: is style imitation possible in principle? To address this question we need to recall that the models are performing style transformation on a sentence base, rather that taking into account the entire corpora. Even if individual sentence cannot be changed to be classified as a different author, it is still possible to modify the entire text to be classified as a targeted author. BT provides some curious examples of transformations, where a "husband" in a female sentence is substituted with a "wife" when transferring to the male style. Although, it might look like this is an appropriate transformation, the change does not represent proper semantic retainment.

Deep learning in style imitation. Deep learning technologies are perfectly suitable for adversarial training, but as illustrated in this thesis AI technologies alone can not reliably retain semantics in the text style imitation. Deep learning technologies are able to deceive deep learning classifiers, but fail to retain original content. Primary reason is lack of reliable mechanisms of enforcing semantic retainment. A^{4}NT supports several type of loss penalties, however they function similar to max edit distance. Max edit distance restricts number of changes in the sentence, but does not check whether the changes are semantically appropriate. Therefore incorporating techniques from different fields would likely result in better style imitation. Absence of correct paraphrasing can be addressed by employing automatic paraphrasing. Modern deep learning techniques were able to successfully perform paraphrasing of diverse data [17, 45]. Even though, these approaches are not directly applicable for style imitation tasks, they should not be neglected. A potential imitation technique would be producing set of possible grammatically correct paraphrases of the original sentence and then choosing the one that brings the authorship classifier closer to a targeted author.

We conclude, that the state-of-the-art models are not able to perform proper transformations even when provided with vast training corpora and simple and grammatically correct sentences.

Self-evaluation. Lack of computational resources had a detrimental effect on the training process of the models. For big size corpora (BG and YG) we reduced the vocabulary size and used the "weaker" set of hyperparameters. The biggest issue was training the GAN for YG. Due to the usage of a cycle loss, and a GAN being two competing networks, we needed to drastically reduce the size of the training dataset. Given unlimited time and computational resources, we would be able to use the entire training datasets, and a bigger vocabulary size for every conducted experiment. We also would be able to establish the

unique set of hyperparameters that would yield the best results for each model. In most case we had enough resources to run the training process only once. However, by using different combinations of learning rate, learning rate decay, max sequence length of the training sentences, dimensions of NN, and embedding size we would likely improve the result for every trained entity. We could also potentially improve the results for A^4NT model by experimenting with different combinations of losses. A^4NT supports cycle loss, and language loss. Given unlimited time, we could have trained and tested whether it is possible to improve the semantic retainment of the transformations. A^4NT also supports both character and word based models. In theory, character-based models might perform better for cases of limited training corpora [36]. We could verify/refute this claim by training both models for every transformation.

Furthermore, we believe that it is possible to improve some of the conducted experiments. For comparison of the imitation success, we employed state-of-the-art deep learning methods, that solve classification tasks successfully. However, we could improve our semantic retainment tests. Although, METEOR and BLEU were illustrated to have a high correlation to human judgement, those metrics can often misjudge a good paraphrase and vice-versa. BLUE does not take into account paraphrases, whereas METEOR is limited by WordNet[1] and internal paraphrase tables. Moreover, the sample size of conducted manual-evaluation is relatively low, compared to the size of tested corpora. We could conduct an extensive user study to address these issues.

Having more time and hardware resource would likely positively reflect on some of the obtained results. However, as mentioned above, the fundamental issue with the deep learning models is generating text instead of "imitationally" paraphrasing the original. Hence, the obtained patterns for the imitation models are likely to remain the same even with the best training possible.

We conducted our research based on the individual author imitation and gender transferring. However, BG also contains the age attribute for every blog author. Due to lack of time we were not able to conduct full scale experiments on age transferring. While BG has an age attribute for every sentence, this attribute is absent in Yelp reviews. Therefore, age imitation results can not be compared across different datasets. Nevertheless, the age imitation can find its way into real-life application.

[1] https://wordnet.princeton.edu/

Chapter 6

Related work

Machine translation. Style imitation can be related to the task of machine translation. Both are based on mapping an input sequence to an output sequence. Both processes are based on technology called sequence-to-sequence networks [14, 40]. Such an entity consists of two networks: (i) an encoder, a recurrent neural network(RNN) that processes an input sentence that maps the input into a latent-variable, (ii) a decoder, a RNN that generates an output sentence based on this latent-variable. NMT models have been illustrated to successfully perform machine translation [7, 19, 34, 43]. Machine translation uses a large amount of paired data for the sentence transformation.

Adversarial training. Recent advances in adversarial training have many implications in style imitation. *Adversarial training* is as a process of training the model on adversarial examples [41]. *Adversarial examples* are possible inputs to a deep learning model that lead to incorrect outputs [14]. Szegedy et al. [41] demonstrates that even NNs that operate at a human level accuracy can be fooled with a probability close to 100% into classifying images close to each other for a human observer as different entities. This results are achieved by adding tiny amount of perturbation to the input image. A famous example of this is provided by an image depicting a panda being classified as "gibbon" after noise injection invisible to a human observer. [33]

Automatic paraphrasing. An imitation model that cannot uphold reliable semantic retention is not be applicable for real-life tasks, regardless of its success in deceiving authorship classifiers. One of the ways to perform style imitation and preserve the original meaning of the corpora would be to produce proper paraphrase of the original text. Resent studies in *Paraphrase generation* [17, 21] have demonstrated success in performing automatic paraphrasing. Xu et al. recruited automatic paraphrasing to successfully transform Shakespeare style into modern English. Naturally, such transformation has similarities with text style imitation. Furthermore, *automatic text simplification* can be considered as transformation of writing style. Automatic text simplification belongs to the field of paraphrase generation and has been studied in a number of papers [24, 37, 38, 44].

Chapter 7

Conclusions

All models studied in this thesis employ recent deep learning techniques and are open-source. The conducted experiments adopt set of the state-of-the-art authorship classifiers to asses imitation success. For accessing the retainment of semantic content we used combination of textual similarity measures and extensive manual evaluation.

To conclude the thesis, the following three questions regarding writing style imitation need to be answered:

Q1 Can the style imitation models successfully perform text style imitation?

Q2 How well can the style imitation models preserve original semantic content?

Q3 How do complexity, and grammar of corpora affect the imitation success?

Turning to Q1, experiment results obtained in Chapter 4 suggest that style-imitation models are partly successful in imitation the target style. We demonstrated that studied methods are able to fool the state of the art deep learning authorship classifiers for balanced datasets. However, the models fail to consistently perform imitation across imbalanced datasets.

Based on the manual and automatic evaluation of semantic retainment conducted in Chapter 3, Q2 was answered negatively. None of the state-of-the-art deep learning style imitation models are able to preserve semantic of the original text. Furthermore all models failed for every tested datasets, meaning that those models can not be used for real-life imitation. Furthermore, manual evaluations conducted in Chapter 4 illustrate that none of the imitation models are not able to consistently produce error-free transformations. Out of studied 100 sentences BG transformations are semantically correct in at most 51%. For YG transformations this number is 21%. Finally, in AB and TO correct transformations are occurring in 16% and less than 5% sentences respectively. The A^4NT model shows the best semantic preservation across all 4 datasets.

Turning to Q3, complexity and grammar were shown to have a substantial affect on imitation success. Style imitation models maintain better results in the authorship classification in cases of "clean" data. Sentence's length has been shown to affect the semantic retainment. Manual evaluation indicates that transformations of shorter sentences more often preserve original meaning. It is clear that majority of changes that the models are introducing are semantically incorrect. Therefore, shorter sentence has a higher chance to stay unchanged and preserve semantics.

Summarizing our comparison of modern style imitation models, it is evident that using only state-of-the-art deep learning techniques is not enough to solve real-life style imitation tasks. We believe that in order to successfully perform style transfer different approaches have to be employed. Incorporating methods from different fields would likely provide better imitation results. In future works, We intend to study different ways of improving existing

techniques of style imitation, we also plan to conduct age imitation experiments.

Bibliography

[1] Pytorch framework. [online]. `https://pytorch.org/`. Accessed: 2019-01-15.

[2] Tensorflow framework. [online]. `https://www.tensorflow.org/`. Accessed: 2019-01-15.

[3] Vice president pence denies he's the 'lodestar' behind anonymous new york times op-ed. [online]. `http://time.com/5388483/mike-pence-denies-writing-anonymous-op-ed/`. Accessed: 2019-01-15.

[4] Afroz, S., Caliskan-Islam, A., Stolerman, A., Greenstadt, R., and McCoy, D. Doppelgänger finder: Taking stylometry to the underground. In *Proceedings of the 2014 IEEE Symposium on Security and Privacy* (2014), pp. 212–226.

[5] Almishari, M., Oguz, E., and Tsudik, G. Fighting Authorship Linkability with Crowdsourcing. In *Proceedings of the second ACM conference on Online social networks* (2014), pp. 69–82.

[6] Argamon, S., Koppel, M., Pennebaker, J. W., and Schler, J. Automatically profiling the author of an anonymous text. In *Communications of the ACM*.

[7] Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations (ICLR)* (2014).

[8] Bennett, D. A 'Gay Girl in Damascus', the Mirage of the 'Authentic Voice' - and the Future of Journalism. In *Mirage in the Desert? Reporting the Arab Spring*, R. L. Keeble and J. Mair, Eds. Abramis, Bury St. Edmunds, 2011, pp. 187–195.

[9] Brennan, M., Afroz, S., and Greenstadt, R. Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity. *ACM Transactions on Information and System Security 15*, 3 (2011).

[10] Coulthard, M. Author identification, idiolect and linguistic uniqueness. *Applied Linguistics 25*, 4 (2004), 431–447.

[11] Denkowski, M., and Lavie, A. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation. ACL* (2014).

[12] Galán-García, P., de la Puerta, J. G., Gómez, C. L., Santos, I., and Bringas, P. G. Supervised machine learning for the detection of troll profiles in Twitter social network: Application to a real case of cyberbullying. In *International Joint Conference of Advances in Intelligent Systems and Computing* (2014), pp. 419–428.

[13] Gatys, L. A., Ecker, A. S., and Bethge, M. Image style transfer using convolutional neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2016).

[14] Goodfellow, I., Bengio, Y., and Courville, A. *Deep learning*. The MIT Press, 2016.

[15] Gröndahl, T., and Asokan, N. Text Analysis in Adversarial Settings: Does Deception Leave a Stylistic Trace? *to appear in CSUR* (Feb 2019), arXiv:1902.08939.

[16] Juola, P. Stylometry and immigration: A case study. *Journal of Law and Policy 21*, 2 (2013), 287–298.

[17] Li, Z., Jiang, X., Shang, L., and Li, H. Paraphrase Generation with Deep Reinforcement Learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2018), pp. 3865–3878.

[18] Liang, B., Li, H., Su, M., Bian, P., Li, X., and Shi, W. Deep Text Classification Can be Fooled. *arXiv e-prints* (Apr 2017), arXiv:1704.08006.

[19] Luong, M.-T., Pham, H., and Manning, C. D. Effective approaches to attention-based neural machine translation. In *Empirical Methods in Natural Language Processing (EMNLP)* (2015), pp. 1412–1421.

[20] Mack, N., Bowers, J., Williams, H., Dozier, G., and Shelton, J. The Best Way to a Strong Defense is a Strong Offense: Mitigating Deanonymization Attacks via Iterative Language Translation. *International Journal of Machine Learning and Computing 5*, 5 (2015), 409–413.

[21] Madnani, N., and Dorr, B. Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Journal of Computational Linguistics 36*, 3 (2010), 341–387.

[22] Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations* (2014), pp. 55–60.

[23] McMenamin, G. R., and Choi, D. *Forensic Linguistics: Advances in Forensic Stylistics*. CRC Press, London, 2002.

[24] Narayan, S., and Gardent, C. Hybrid simplification using deep semantics and machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)* (2014), pp. 435–445.

[25] Narayanan, A., Paskov, H., Gong, N. Z., Bethencourt, J., Stefanov, E., Shin, E. C. R., and Song, D. On the feasibility of internet-scale author identification. In *Proceedings of the 2012 IEEE Symposium on Security and Privacy* (2012), pp. 300–314.

[26] Neal, T., Sundararajan, K., Fatima, A., Yan, Y., Xiang, Y., and Woodard, D. Surveying stylometry techniques and applications. *ACM Computing Surveys 50*, 6 (2017), 86:1–86:36.

[27] Overdorf, R., and Greenstadt, R. Blogs, twitter feeds, and reddit comments: Cross-domain authorship attribution.

[28] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)* (Philadelphia, 2002), pp. 311–318.

[29] Prabhumoye, S., Tsvetkov, Y., Salakhutdinov, R., and Black, A. W. Style Transfer Through Back-Translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)* (2018), pp. 866–876.

[30] Reddy, S., and Knight, K. Obfuscating gender in social media writing. In *Proc. of Workshop on Natural Language Processing and Computational Social Science* (2016), pp. 17–26.

[31] Samanta, S., and Mehta, S. Towards Crafting Text Adversarial Samples. *arXiv e-prints* (Jul 2017), arXiv:1707.02812.

[32] Schler, J., Koppel, M., Argamon, S., and Pennebaker, J. Effects of age and gender on blogging. In *AAAI spring symposium: Computational approaches to analyzing weblogs* (2006).

[33] Schler, J., Koppel, M., Argamon, S., and Pennebaker, J. Explaining and harnessing adversarial examples. In *ICLR* (2015).

[34] Sennrich, R., Haddow, B., and Birch, A. Controlling politeness in neural machine translation via side constraints. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)* (2016), pp. 35–40.

[35] Shen, T., Lei, T., Barzilay, R., and Jaakkola, T. Style transfer from non-parallel text by cross-alignment. In *Proceedings of Neural Information Processing Systems NIPS* (2017).

[36] Shetty, R., Schiele, B., and Fritz, M. A^{4}nt: Author attribute anonymity by adversarial training of neural machine translation. In *27th USENIX Security Symposium (USENIX Security 18)* (2018), pp. 1633–1650.

[37] Siddharthan, A. Complex lexico-syntactic reformulation of sentences using typed dependency representations. In *Proceedings of the 6th International Natural Language Generation Conference* (2010), pp. 125–133.

[38] Siddharthan, A. Text Simplification using Typed Dependencies: A Comparision of the Robustness of Different Generation Strategies. In *Proceedings of the 13th European Workshop on Natural Language Generation* (2011), pp. 2–11.

[39] Solorio, T., Hasan, R., and Mizan, M. A Case Study of Sockpuppet Detection in Wikipedia. In *Proceedings of the Workshop on Language in Social Media* (2013), pp. 59–68.

[40] Sutskever, I., Vinyals, O., and Le, Q. V. Sequence to sequence learning with neural networks. In *Advances in neural information* (2014).

[41] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *arXiv e-prints* (Dec 2013), arXiv:1312.6199.

[42] Woolley, J. T., and Peters., G. The american presidency project. [online]. `http://www.presidency.ucsb.edu`, 1999. Accessed: 2019-01-15.

[43] Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Åukasz Kaiser, Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., and Dean, J. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR abs/1609.08144* (2016).

[44] Wubben, S., van den Bosch, A., and Krahmer, E. Sentence simplification by monolingual machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)* (2012), pp. 1015–1024.

[45] Xu, W., Ritter, A., Dolan, B., Grishman, R., and Cherry, C. Paraphrasing for style. In *Proceedings of COLING* (2012), pp. 2899–2914.

Appendix A

Used hyperparameters

A.1 The YG CAE model trained by Shrimai Prabhumoye

In this section the CAE YG model trained by Prabhumoye is described. Table A.1 indicates that every transformation is represented by a small set of repeating words. In the absence of any semantic enforcement mechanism this CAE model was trained to reach substantially high imitation performance at the expense of semantic retainment.

Original sentence	**Transformed sentence**
all excellent and perfectly cooked .	ice-cream muffins teriyaki pescatarian together
i could swim in their gin and tonics .	ice-cream postino patrick evans recurring together
family propane has earned my business .	deluxe kept knobs caught together
decent price for some good burgers and milkshake .	deluxe postino patrick evans recurring together
a really cool and upscale place .	ice-cream muffins teriyaki them xx together
he was talkative , he was cool .	ice-cream jacks target default teriyaki pescatarian together
chicken chow mien was too soupy and just okay .	deluxe jacks teriyaki caught handcrafted rito ice-cream muffins spoons together
the greens are amazing !	mtv stripburger kept teriyaki together;

Table A.1: Transformation examples from the CAE model used in BT paper

A.2 Used hyperparameters

In Table A.2 hyperparameters for training the representative models across 4 datasets are presented. The hyperparameters that are not presented were kept default. The code for the original models can be found on the following github repositories [1].

[1] https://github.com/shentianxiao/language-style-transfer, https://github.com/shrimai/Style-Transfer-Through-Back-Translation, https://github.com/rakshithShetty/A4NT-author-masking

	BG		YG		AB	
CAE	Gen	dim_emb: 100 dim_y: 200 dim_z: 500 n_layers: 1 l_rate: 0.0005 max_seq_len: 20 batch_size: 32	Gen	dim_emb: 300 dim_y: 200 dim_z: 500 n_layers: 1 l_rate: 0.0005 max_seq_len: 20 batch_size: 32	Gen	dim_emb: 300 dim_y: 200 dim_z: 500 n_layers: 1 l_rate: 0.0005 max_seq_len: 20 batch_size: 64
BT	Gen_{male}	optim: sgd word_vec_size: 300 rnn_size: 500 layers: 2 l_rate: 1.0 l_rate_decay: 0.5 max_seq_len: 50 batch_size: 32	Gen_{male}	optim: sgd word_vec_size: 300 rnn_size: 500 layers: 2 l_rate: 1.0 l_rate_decay: 0.5 max_seq_len: 50 batch_size: Pre-Trained	Gen_{bob}	optim: sgd word_vec_size: 300 rnn_size: 500 layers: 2 l_rate: 1.0 l_rate_decay: 0.5 max_seq_len: 50 batch_size: 64
	Gen_{fem}	optim: sgd word_vec_size: 300 rnn_size: 500 layers: 2 l_rate: 1.0 l_rate_decay: 0.5 max_seq_len: 50 batch_size: 32	Gen_{fem}	optim: sgd word_vec_size: 300 rnn_size: 500 layers: 2 l_rate: 1.0 l_rate_decay: 0.5 max_seq_len: 50 batch_size: Pre-Trained	Gen_{alice}	optim: sgd word_vec_size: 300 rnn_size: 500 layers: 2 l_rate: 1.0 l_rate_decay: 0.5 max_seq_len: 50 batch_size: 64
	CNN	l_rate: 1.0 l_rate_decay: 0.5 max_seq_len: 50 optim: sgd batch_size: 64	CNN	l_rate: 1.0 l_rate_decay: 0.5 max_seq_len: 50 optim: sgd batch_size: 64	CNN	l_rate: 1.0 l_rate_decay: 0.5 max_seq_len: 50 optim: sgd batch_size: 64
A^4NT	Gen	l_rate_gen: 0.001 max_seq_len: 50 batch_size: 32 atoms: word emb_size: 512 enc_hid_depth:1 enc_hid_size:512 dec_hid_depth:1 dec_hid_size:512	Gen	l_rate_gen: 0.001 max_seq_len: 50 batch_size: 32 atoms: word emb_size: 512 enc_hid_depth:1 enc_hid_size:512 dec_hid_depth:1 dec_hid_size:512	Gen	l_rate_gen: 0.001 max_seq_len: 50 batch_size: 32 atoms: word emb_size: 512 enc_hid_depth:1 enc_hid_size:512 dec_hid_depth:1 dec_hid_size:512
	GAN	cycle_loss_w: Pre-Trained cycle_loss_type: Pre-Trained l_rate_gen: 0.0001 l_rate_eval: 0.001 max_seq_len: 50 batch_size: Pre-Trained atoms: word emb_size: 512 enc_hid_depth:1 enc_hid_size:512 dec_hid_depth:1 dec_hid_size:512	GAN	cycle_loss_w: 1 cycle_loss_type: enc l_rate_gen: 0.0001 l_rate_eval: 0.001 max_seq_len: 50 batch_size: 32 atoms: word emb_size: 512 enc_hid_depth:1 enc_hid_size:512 dec_hid_depth:1 dec_hid_size:512	GAN	cycle_loss_w: 1 cycle_loss_type: enc l_rate_gen: 0.0001 l_rate_eval: 0.001 max_seq_len: 50 batch_size: 32 atoms: word emb_size: 512 enc_hid_depth:1 enc_hid_size:512 dec_hid_depth:1 dec_hid_size:512
	LSTM	l_rate_eval: 0.001 max_seq_len: 50 batch_size: 128 atoms: word emb_size: 512 hid_depth:1 hid_size:512	$LSTM^1$	l_rate_eval: 0.0001 max_seq_len: 50 batch_size: 16 atoms: word emb_size: 512 hid_depth:1 hid_size:512	LSTM	l_rate_eval: 0.0001 max_seq_len: 50 batch_size: 16 atoms: word emb_size: 512 hid_depth:1 hid_size:512
			$LSTM^2$	l_rate_eval: 0.0001 max_seq_len: 50 batch_size: 16 atoms: word emb_size: 512 hid_depth:1 hid_size:512		

Table A.2: Used hyperparameters.